



**An Introduction to Multivariate
Statistical Analysis**
多元统计分析导论

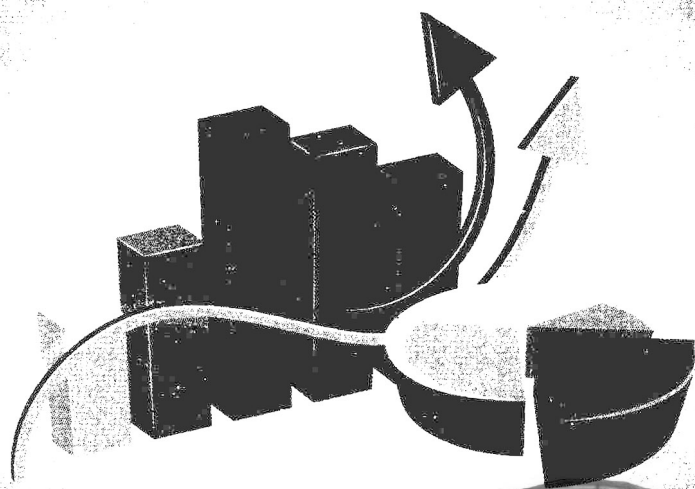
(第3版)



人民邮电出版社
POSTS & TELECOM PRESS

TURING

图灵数学·统计学丛书 51



An Introduction to Multivariate Statistical Analysis

多元统计分析导论

(第3版)

人民邮电出版社
北京

图书在版编目(CIP)数据

多元统计分析导论/(美)安德森(Anderson, T. W.)
著;张润楚等译. —北京:人民邮电出版社, 2010.12
(图灵数学·统计学丛书)
书名原文: An Introduction to Multivariate
Statistical Analysis, Third Edition
ISBN 978-7-115-24118-4

I. ①多… II. ①安… ②张… III. ①多元分
析: 统计分析 IV. ①0212.4

中国版本图书馆CIP数据核字(2010)第202633号

内 容 提 要

本书是世界知名统计学家的力作,主要内容有多元正态分布、方差分析、回归分析、因子分析、椭圆等高分布、相依性模式、图模型。附录中还列出了矩阵理论、Wilk 似然准则和其他常用检验的显著性水平的分位数。

本书在世界各高等学校中广为采用,是一本经典的多元统计分析课程的教材,也可供相关统计研究人员、应用多元统计的科技工作者参考。

图灵数学·统计学丛书

多元统计分析导论(第3版)

-
- ◆ 著 [美] T. W. Anderson
 - 译 张润楚 程轶 等
 - 责任编辑 明永玲
 - ◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街14号
邮编 100061 电子邮件 315@ptpress.com.cn
网址: <http://www.ptpress.com.cn>
北京隆昌伟业印刷有限公司印刷
 - ◆ 开本: 700×1000 1/16
印张: 35
字数: 703千字 2010年12月第1版
印数: 1~2500册 2010年12月北京第1次印刷
- 著作权合同登记号 图字: 01-2007-5300号

ISBN 978-7-115-24118-4

定价: 89.00元

读者服务热线: (010)51095186 印装质量热线: (010) 67129223

反盗版热线: (010)67171154

版 权 声 明

Original edition, entitled *An Introduction to Multivariate Statistical Analysis, Third Edition*, by T. W. Anderson, ISBN 0-471-36091-0, published by John Wiley & Sons, Inc.

Copyright © 2003 John Wiley & Sons, Inc. All rights reserved. This translation published under License.

Simplified Chinese translation edition published by POSTS & TELECOM PRESS
Copyright © 2010.

Copies of this book sold without a Wiley sticker on the cover are unauthorized and illegal.

本书简体字中文版由 John Wiley & Sons, Inc. 授权人民邮电出版社独家出版并在全世界销售. 未经出版者书面许可, 不得以任何方式复制本书内容.

本书封底贴有 John Wiley & Sons, Inc. 激光防伪标签, 无标签者不得销售.
版权所有, 侵权必究.

译者序

作为人类认知世界最基础的共用科学之一的统计学,像哲学、数学(主要指纯粹数学)一样,应用非常广泛.与哲学和数学所不同的是,哲学通过对世界的总体看法和思考来认知世界,数学通过抽象地研究数量、图形、符号等形式逻辑理论和方法来认知世界,而统计学则通过对世界观察得到各种形态的信息资料(称为数据)进行分析,研究有关的推断理论和方法来认知世界的.哲学、数学和统计学的发展各自相对独立,又相互依赖和支撑.从历史上看,统计学、数学和哲学思想的形成应该是相互交替进行的.但归根结底,人类的认知应该首先来源于对世界的观察.由此看来,通过观察进行归纳和推断的统计思想应该更为基本.哲学和数学是对观察结果的抽象,它们经过发展所形成的各种形式逻辑和辩证逻辑使得推理更为严密,因而反过来又会促进更为科学地观察和分析推断世界.这点也许是为什么统计学思想的产生和发展虽然很早但现代统计学的形成却比较晚的原因之一.

我们有理由把统计学定义为一门通用科学,它研究如何有效地观察世界,得到信息(各种形态的数据资料),并由此来认知世界,研究的问题包括如何收集、整理和分析数据以及进行推断.统计学离不开对世界的观察,也离不开数据和对数据的分析.世界是复杂的,存在各式各样的研究对象和各式各样的研究问题.如果我们对任何一个研究对象的某个研究问题加以分析,大都可以归结为对变量之间关系的研究.因此通过得到的数据研究变量之间的关系,就成为普遍关心的问题.于是,在统计学中,多元统计分析就自然地成为最重要的研究分支之一.

对多元统计方法开始系统研究可以追溯到20世纪20年代甚至更早,本书的作者T. W. Anderson,还有我国的许宝禄先生,就是研究这类方法的先驱者.本书的第1版出版于1958年,它是对成书之前形成的各种多元统计理论和方法的最早的总结.由于其理论的系统性和深刻性以及方法的多样性,使其成为这个分支的第一个经典之作.多元统计的理论和方法还在不断发展.正如作者在本书的第2版和第3版的前言中所说,在第1版之后,多元统计理论和方法的许多方面都得到很快的发展.现在,国际上在发表这方面研究论文的同时,出版了大量的多元统计分析的著作和教材,许多都可以用来作为研究和应用的参考,但我们认为,这本由T. W. Anderson写的书仍然是我们想了解这个领域的首选之作,尤其是在了解理论和方法方面,它可以帮助我们了解对问题的开始提法和研究方向.

现在,多元统计方法已经被广泛地用于各个领域,包括金融学、经济学、社会学、生物学、医学、物理、化学、工学和农学等各学科,还有环保、气象等各个领域.

希望这本书的中文版能促进我国在这个领域理论和应用的发展.

本书的翻译是由张润楚、程轶、庞芳、薛晓琳和曹琼合作完成的, 张润楚翻译了第 15 章和附录, 程轶、庞芳、薛晓琳和曹琼分别翻译了第 1~5 章、第 6~8 章、第 9~11 章、第 12~14 章, 最后由程轶和张润楚进行审阅和协调统一. 郑贵鹏也参加了部分翻译工作并对书稿编辑和形成 CTX 文件做了不少工作, 图灵公司编辑们为本书的出版给予了大力支持, 包括仔细审校书稿, 我们特此表示感谢. 由于译者的学识水平以及时间所限, 本书的译文肯定会存在缺点甚至错误之处, 肯请读者给予批评指正.

张润楚

于南开大学和东北师范大学

2010 年 4 月

第 3 版前言

本书的前两版用来教授学生学习多元统计分析理论和方法已经快 40 年了. 同时该书在帮助广大统计学者增强对这个领域的了解和熟悉有关知识方面也发挥了作用. 自第二版出版以来, 多元分析在许多方向上已经得到很大发展和开拓. 在这个新版本里, 我不想论述所有扩大了的范围, 而只是选择对理论方法和综合理解特别有用的若干方面进行阐述.

前两个版本包含了只要用简单的计算器就可以完成的若干方法. 然而, 在 21 世纪的今天, 计算技术已经得到高度发展, 不可能用一本一般的数学理论书来包括所有的计算方法. 统计学某些方面也用到了再抽样技术这样的计算手段, 但这些并没有在这一版本里介绍.

多元统计处理相互间具有联系的变量. 这个版本我们仍用几章来讲述相关性的度量和独立性检验. 我们加进了新的一章, “相依性模式, 图模型”. 所谓图模型, 是由一组顶点 (也叫结点) 连同一组连线 (也叫边) 构成的图, 这些顶点用于识别观测到的变量, 而这些连线用于反映变量间相依性关系. 这样, 图的代数理论便是路径分析和因果链研究的一个自然发展和产物. 一个图可以表达依时间先后或逻辑关系形成的序列, 也可以设想为一组变量是导致另一组变量变化的原因.

本版中, 我们系统介绍的一个新的论题是椭球等高分布. 我们知道, 由均值向量和协方差阵刻画的多元正态分布具有一个局限, 即变量的四阶矩由一阶矩和二阶矩就可以决定. 而椭球等高分布类放宽了这个限制, 在这个类中的分布密度也具有像正态密度一样的椭球型等高密度, 但是其变量的四阶矩多出一个自由度. 这个论题在书中新增的有关章节里给予详细阐述.

我们在第 12 章和第 13 章中详尽阐述的降秩回归提供了一个方法, 用以减少在一组变量对另外一组变量的回归中要估计的回归系数个数. 这一方法包含着在一个联立方程模型中一个方程的有限信息极大似然估计.

第 3 版的写作得益于前两版读者的建议和评论, 还有这一版本审稿人的意见. 除了前两个版本前言列出的读者以外, 我还要感谢 Michael Perlman 和 Kathy Richards, 他们为完成这个稿子给予了很大帮助.

T. W. Anderson

2003 年 2 月于加州斯坦福

第 2 版前言

自本书的第 1 版出版已经过去 26 年了. 在这段时间里, 多元统计分析得到很大发展, 特别在本书第 1 版所提到的那些方面. 这个新版本的目的是通过大量修订、改写和增添, 让这本书涵盖这个领域最新的发展. 这个版本仍保留原来的基本途径, 即对于每个主题由几种测量或刻画以及对它们性质的研究组成的统计方法在数学上的严格发展. 论题的一般梗概仍然保留.

在这个版本中, 从其他方面讨论了极大似然方法. 在均值向量和协方差阵的点估计中, 还介绍了关于某些损失函数比极大似然估计更好的估计, 如 Stein 估计和贝叶斯估计. 在假设检验方面, 似然比检验通过其他不变性方法处理得以补充. 给出了一些关于分布和渐近分布的新结果, 列表给出某些分位数. 对这些方法的性质, 比如功效函数、容许性、无偏性以及功效函数的单调性, 都进行了研究. 对均值和协方差的联立置信区间进行了研究. 用因子分析的一章代替了第 1 版中勾画零碎结果的那一章. 还介绍了一些新的论题, 包括联立方程模型和线性函数关系. 附加问题中介绍了进一步的结果.

本书不可能覆盖所有的有关材料, 但应该已经包括了那些我认为是最重要的内容. 关于直到 1966 年的论文和直到 1970 年的书的综合参考文献, 读者可以参看由 Anderson, Das Gupta 和 Styan (1972) 所著的 *A Bibliography of Multivariate Statistical Analysis*. 进一步的参考文献可以在 Subrahmaniam 和 Subrahmaniam (1973) 著的 *Multivariate Analysis: A Selected and Abstracted Bibliography, 1957-1972* 中查到.

我要感谢许多学生、同事和朋友, 他们提出了建议和帮助. 他们包括 Yasuo Amemiya, James Berger, Byoung-Seon Choi, Arthur Cohen, Margery Cruise, Somesh Das Gupta, Kai-Tai Fang, Gene Golub, Aaron Han, Takeshi Hayakawa, Jogi Henna, Huang Hsu, Fred Huffer, Mituaki Huzii, Jack Kiefer, Mark Knowles, Sue Leurgans, Alex McMillan, Masashi No, Ingram Olkin, Kartik Patel, Michael Perlman, Allen Sampson, Ashis Sen Gupta, Andrew Siegel, Charles Stein, Patrick Strout, Akimichi Takemura, Joe Verducci, Marlos Viana 和 Y. Yajima. 在本书写作过程中我还得到 Dorothy Anderson, Alice Lundin, Amy Schwartz 和 Pat Struse 的帮助. 特别感谢 Johanne Thiffault 和 George P. H. Styan 的关心. 本书的出版得到 Army Research Office, National Science Foundation, the Office of Naval Research 和 IBM Systems Research Institute 的资助.

附录 B 给出了 7 个分位数表, 以供进行检验方法使用. 表 1、表 5 和表 7 分别是 E. S. Pearson 和 H. O. Hartley 的著作 *Biometrika Tables for Statisticians* 第二卷的表 47、表 50 和表 53; 表 2 是根据由 A. W. Davis 准备的和发表在 *Biometrika* (1970a), *Annals of the Institute of Statistical Mathematics* (1970b) 和 *Communications in Statistics, B. Simulation and Computation* (1980) 的 3 个表做出来的. 表 3 和表 4 分别是由 Ziro Yamauti (1977) 编辑和由 Japanese Standards Association 发表的 *Concise Statistical Tables* 中的表 6.3 和表 6.4, 那是 *Statistical Tables and Formulas with Computer Applications, JSA-1972* 的一个简明版本. 表 6 是 B. N. Nagarsenker 和 K. C. S. Pillai 以及 Aerospace Research Laboratories (1972) 发表的 *The Distribution the Sphericity Test Criterion*, ARL 72-0154 的表 3. 感谢上述作者和出版者允许我在本书中使用这些表.

T. W. Anderson

1984 年 6 月于加州斯坦福

第 1 版前言

本书的主要目标是成为多元统计两学期课程的教材。同时希望该书能作为那些不是学生的统计学者了解这个领域各方面的一个导引,其他统计学家也可以把它用做参考书。

这本书曾以讲义笔记的形式在哥伦比亚大学的一门两学期研究生系列课中使用多年;前 6 章构成第一学期的教学内容,重点放在相关性理论方面。假设读者熟悉一元统计的一般理论,特别是了解基于一元正态分布的方法。矩阵代数的知识也是必须具备的;不过,关于这方面的内容在附录中已经介绍了。

希望更基础和更重要的论题在这里得到处理,虽然在某种程度上这种涵盖与个人喜好有关。某些最近的前沿发展只是在后面的章节有简短的介绍。

本书在很大的程度上使用极大似然方法。这引导出许多合理的结果,且在某些情况下可以证明这些结果是最优的。然而,在许多情况下,这些期望的或最优的结果还有缺陷。

在这本书稿不断完善的几年里,大量的学生和同事给予了很大的帮助。特别要提到的是 Allan Birnbaum, Harold Hotelling, Jacob Horowitz, Howard Levene, Ingram Olkin, Gobind Seth, Charles Stein 和 Henry Teicher。还要感谢哥伦比亚大学的研究生数学统计学会在从讲义笔记到形成这本书的过程中所给予的协助。本书的写作得到了 Office of Naval Research 的资助,在这里一并致谢。

T. W. Anderson

1957 年 12 月于加州斯坦福

目 录

第 1 章 引论	1	5.2 广义 T^2 统计量的推导及分布	124
1.1 多元统计分析	1	5.3 T^2 统计量的应用	129
1.2 多元正态分布	2	5.4 备择假设下 T^2 的分布, 功效函数	135
第 2 章 多元正态分布	5	5.5 协方差阵不等时的两样本问题	136
2.1 引言	5	5.6 T^2 检验的一些最优性质	139
2.2 多元分布的概念	5	5.7 椭球等高分布	146
2.3 多元正态分布	10	习题	147
2.4 正态分布变量线性组合的分布, 变量的独立性, 边缘分布	17	第 6 章 观察值的分类	151
2.5 条件分布和多重相关系数	24	6.1 分类问题	151
2.6 特征函数和矩	29	6.2 精确分类的标准	151
2.7 椭球等高分布	33	6.3 概率分布已知的两总体的判别	154
习题	40	6.4 两多元正态总体的判别	157
第 3 章 均值向量和协方差阵的估计	47	6.5 具有估计参数的两多元正态总体的判别	160
3.1 引言	47	6.6 误判概率	165
3.2 均值向量和协方差阵的极大似然估计	47	6.7 多总体的分类	170
3.3 样本均值向量的分布, 协方差阵已知时均值的推断	53	6.8 多个多元正态总体的分类	173
3.4 均值向量的估计的理论性质	59	6.9 多个多元正态总体分类的一个例子	175
3.5 均值的改良估计	66	6.10 具有不同协方差阵的两多元正态总体的分类	177
3.6 椭球等高分布	73	习题	182
习题	78	第 7 章 样本协方差阵和样本广义方差的分布	184
第 4 章 样本相关系数的分布和利用	84	7.1 引言	184
4.1 引言	84	7.2 Wishart 分布	184
4.2 二元变量样本的相关系数	85	7.3 Wishart 分布的一些性质	189
4.3 偏相关系数, 条件分布	98	7.4 Cochran 定理	192
4.4 多重相关系数	104		
4.5 椭球等高分布	114		
习题	118		
第 5 章 广义 T^2 统计量	124		
5.1 引言	124		

7.5	广义方差	194	9.8	两个变量集的情形	298
7.6	总体协方差阵为对角矩阵时 相关系数集的分布	198	9.9	似然比检验的容许性	301
7.7	逆 Wishart 分布, 协方差阵的 贝叶斯估计	200	9.10	子集间独立性检验的功效函 数的单调性	302
7.8	协方差阵的改进估计	203	9.11	椭球等高分布	304
7.9	椭球等高分布	208	习题		307
习题		210	第 10 章	协方差阵相等以及均值向量 和协方差阵均相等的假设 检验	309
第 8 章	一般的线性假设检验, 多元方 差分析	215	10.1	引言	309
8.1	引言	215	10.2	检验几个协方差阵相等的 准则	309
8.2	多元线性回归中的参数估计	216	10.3	检验几个正态分布相等的 准则	311
8.3	关于回归系数线性假设检验 的似然比准则	220	10.4	准则的分布	313
8.4	假设成立时似然比准则的 分布	225	10.5	准则的分布的渐近展开	319
8.5	似然比准则的分布的渐近 展开	234	10.6	两个总体的情形	321
8.6	检验线性假设的其他准则	242	10.7	检验协方差阵与给定矩阵成 正比的假设; 球形检验	325
8.7	关于回归系数矩阵和置信区 域的假设检验	251	10.8	检验一个协方差阵等于一个 给定的矩阵的假设	329
8.8	具有相同协方差阵的几个正 态分布均值相等的检验	254	10.9	检验均值向量和协方差阵分 别等于给定的向量和矩阵的 假设	334
8.9	多元方差分析	258	10.10	检验的容许性	336
8.10	检验的一些最优性质	263	10.11	椭球等高分布族	339
8.11	椭球等高分布	276	习题		342
习题		279	第 11 章	主成分	346
第 9 章	检验变量集间的独立性	285	11.1	引言	346
9.1	引言	285	11.2	总体中主成分的定义	347
9.2	变量集独立性检验的似然 比准则	285	11.3	主成分和它们的方差的极大 似然估计	352
9.3	当原假设为真时似然比准则 的分布	289	11.4	主成分的极大似然估计的 计算	353
9.4	似然比准则的分布的渐近 展开	292	11.5	例子	355
9.5	其他准则	293	11.6	统计推断	357
9.6	逐步下降法	294	11.7	关于协方差阵的特征根的 假设检验	360
9.7	例子	297			

11.8 椭球等高分布	363	13.8 椭球等高分布	424
习题	364	习题	427
第 12 章 典型相关和典型变量	367	第 14 章 因子分析	428
12.1 引言	367	14.1 引言	428
12.2 总体的典型相关和典型 变量	368	14.2 模型	428
12.3 典型相关和典型变量的 估计	376	14.3 随机正交因子的极大似然估 计量	433
12.4 统计推断	379	14.4 不变因子的估计	441
12.5 一个例子	381	14.5 因子的解释和变换	442
12.6 线性相关期望值	383	14.6 指定零识别的估计	444
12.7 降秩回归	387	14.7 因子得分的估计	445
12.8 联立方程模型	388	习题	446
习题	396	第 15 章 相依性模式, 图模型	447
第 13 章 特征根和特征向量的分布	398	15.1 引言	447
13.1 引言	398	15.2 无向图	448
13.2 两个 Wishart 矩阵的 情况	398	15.3 有向图	453
13.3 一个非奇异 Wishart 矩 阵的情况	405	15.4 链图	458
13.4 典型相关	409	15.5 统计推断	460
13.5 有一个 Wishart 矩阵情况 下的渐近分布	410	附录 A 矩阵理论	469
13.6 有两个 Wishart 矩阵情况 下的渐近分布	413	A.1 矩阵和矩阵运算的定义	469
13.7 一个回归模型下的渐近 分布	417	A.2 特征根和特征向量	473
		A.3 分块向量和分块矩阵	476
		A.4 其他方面的一些结果	479
		A.5 Gram-Schmidt 正交化和线性 方程组的解	484
		附录 B 表	487
		参考文献	525

第1章 引 论

1.1 多元统计分析

多元统计分析的数据是由在若干个体或对象上的多组测量构成的. 这个样本数据, 可能是从某城市所有在校儿童中随机抽取的一些儿童的身高和体重; 也可能是一个测量集族, 比如两种鸢尾植物花瓣的长度和宽度以及萼片的长度和宽度; 或者又可能是若干学生在一系列心理测试中的得分.

可将一个个体上的多个测量写成一个列向量, 我们把这个向量整体看做是来自某多元总体或分布的一个观测. 当此个体是随机抽取时, 我们认为其测量对应的向量是一个随机向量, 它和总体有相同的分布或概率规则. 样本中所有个体的观测组成了向量样本, 将这些向量排成一行就形成了一个观测矩阵.^① 从而那些需要分析的数据就表示为一个或若干个矩阵.

我们将会看到, 把每个观测向量当做欧氏空间里的一个点有助于直观地表示数据和理解方法, 其中观测向量的每一个坐标相应于一个测量或变量. 实际上, 统计分析首先要做的就是用图表示数据. 大多数统计学家仅限于二维作图, 即将观测向量的两个坐标依次表示在图中.

对于单变量分布, 其重要的特征是作为位置测度的均值和作为变异性测度的标准差. 同样, 对于单变量的样本, 其均值和标准差也是很重要的测度. 在多元分析中, 无论总体还是样本, 不同测量的均值和方差之间具有相应的相关性. 多元分析的一个重要内容就是研究不同变量间的相依性. 两个变量间的相依性可以用它们的协方差表示, 即它们分别与各自均值离差之积的平均. 用相应的标准差将协方差标准化, 就得到了相关系数, 它反映了这两个变量的相依程度. 因此对于多元变量, 可选择均值向量 (由单变量的均值组成) 和协方差阵 (由单变量的方差和两两变量的协方差组成) 作为一组描述统计量. 或者也可以选择均值向量、所有单变量的标准差和相关阵作为一组描述统计量集合. 这两组统计量包含相同的信息. 这些参量描述了一个总体或概率分布的位置、变异性 and 相依性. 多元正态分布就是完全由其均值向量和协方差阵决定的, 因此其样本均值向量和协方差阵就是一组充分统计量.

多元分析的基本内容是刻画和分析变量之间的相依性、变量集之间的相依性

① 当在纸上列表记录数据时, 会很自然地将一个个体的测量记为表上的一行, 因此一个个体相当于一个行向量. 但因为我们更喜欢用列向量进行代数运算, 所以用列向量来表示一个观测. (在实际中, 这些数据也可能记录在卡片、磁带或磁盘上.)

以及变量与变量集之间的相依性. 多重相关系数是相关系数的扩展, 它刻画了一个变量和一组变量之间的关系. 偏相关系数刻画了两个变量在除去其他相关变量影响后的相依程度. 从样本中得到的各种相关系数是用来估计总体分布中相应的相关系数. 本书将讨论独立假设检验, 以及从多元正态分布抽样的估计性质和检验方法.

多元总体的许多统计问题是单变量总体问题的直接类比, 解决这些问题的方法也是相似的. 例如, 在单变量情形下, 我们可能希望检验一个变量的均值为零的假设; 在多元情形下, 我们就可能希望检验若干个变量的均值组成的均值向量为零向量的假设. 前一个假设用到的学生 t 检验的类比是广义 T^2 检验. 单变量的方差分析也适用于观测值向量. 在回归分析中, 因变量可能是一个变量向量. 因此方差的比较就推广为协方差阵的比较.

单变量统计量的检验方法推广到多元情形下需要考虑变量间的相依性, 这些方法是不依赖坐标系的, 即这些方法相对于那些关于原假设不变的线性变换是不变的. 在一些问题中, 可能会有很多族检验是不变的, 这时则需要做出选择. 因此需要考虑检验的最优性质.

然而, 在另一些情况下, 选择一个坐标系是很重要的, 它可使变量具有所希望的统计性质. 有人会说它们具有正态分布或样本的固有特征. 这些和矩阵的典型型的代数问题有很大关联. 例如, 找一组变量的具有极大方差或极小方差的正规线性组合 (找主成分) 就相当于找一个坐标轴的旋转, 使得协方差阵变成对角形式. 另一个例子是刻画两个变量集之间的相依性 (找典型相关). 这些问题涉及不同矩阵的特征根和特征向量, 也要考虑相应样本量的统计性质.

一些统计问题产生于均值和协方差被限定的模型中. 因子分析可能是基于一个 (总体) 协方差阵, 这个协方差阵是一个正定对角矩阵与一个降秩的半正定矩阵之和; 线性结构关系也有相似的形式. 计量经济学中的联立方程组就是一个特殊模型的例子.

1.2 多元正态分布

本书讨论的统计方法是针对多元正态分布的, 但在抽样分布不是正态时, 许多方法也是有效的. 基于正态分布进行统计分析的一个主要原因是, 这个概率模型可以很好地近似很多抽样总体中连续测量的分布. 事实上, 大多数数据统计分析的方法和理论已经发展成熟. 一些数学家, 如 Adrian (1808), Laplace (1811), Plana (1813), Gauss (1823) 和 Bravais (1846) 研究了二元正态密度. 遗传学家 Francis Galton 在研究从一个父代上和一个子代上得到的成对测量时, 引入了相关、回归和同方差性的思想. [例如, 见 Galton (1889).] 他把多元正态分布理论表达为样本观测性质的一种推广.

Karl Pearson 和其他人继续发展了多元正态分布的理论,并在遗传学、生物学和其他领域的研究中使用了不同类型的相关系数.^① R. A. Fisher 进一步发展了应用于农学、植物学和人类学的方法,包括分类问题的判别式函数.在另一个方向上,智力测验得分的分析引出了另一个理论,它包括因子分析,即基于正态分布的抽样理论.统计学家发现,在这些情况下以及在农业实验、工程问题、特殊的经济问题和其他领域里,多元正态分布是足够近似于总体分布的,从而,基于正态模型的统计分析是合理的.

单变量正态分布的出现通常是由于所要研究的效应是若干独立随机效应之和.类似地,多元正态分布的出现通常是由于多重测量是一些较小的独立效应之和.类似于单变量中心极限定理可以导出单变量正态分布,多个变量的广义中心极限定理也可以导出多元正态分布.

基于正态分布的统计理论的优点是,基于此的多元方法已被广泛地发展并可以进行有组织和系统性的研究.这不仅是因为这些方法在实际应用中的需要,还因为正态理论也适应于严格的数学处理.这些恰当的分析方法主要运用了矩阵代数的标准运算;可以得到许多相关的统计量的分布或至少可得到其特征;并且在许多情况下,可以推断方法的最优性质.

本书从多元正态分布的角度陈述所推断的问题,在此基础上研究有效且通常最优的方法,并且计算其显著性水平和置信水平.这个途径具有阐述上的一致性和严密性,但是,就其本质而言,显然它不能详尽地阐述全部多元统计分析的内容.这些方法适用于许多非正态分布,但是它们的充分性还值得怀疑.大致来讲,根据中心极限定理的运算,关于均值的推断是稳健的,但是关于协方差的推断对正态性是敏感的,样本协方差的变化依赖分布的四阶矩.

正态方法关于阶数大于二的矩的不变性可以弱化到一大类椭球等高分布上.在单变量情形中,正态分布可由其均值和方差决定,高阶矩以及峰度和长尾等性质都是均值和方差的函数.同样,在多元情形下,均值向量和协方差阵,或者说,均值向量、方差向量及相关系数阵决定了分布的所有性质.这个限制可以简化为一类很广泛的椭球等高分布.这个类保持了相依性的结构,且允许出现更一般的峰度和长尾.对此的研究可以引出更多稳健方法.

计算机技术的发展给多元统计分析带来了多方面的改革.在单变量统计中,现代计算机支持用再抽样的方法去计算观测的变化和结果的显著性,比如自助法和交叉核实法.这些方法论弱化了对分位数表的依赖,同时也消减了一些正态分布的限制.

非参技术适用于对潜在分布一无所知的情况.由于篇幅所限,本书不涉及这方

^① 关于相关性思想发展的详细研究,见 Walker (1931).

面的内容,也不讨论数据分析的其他问题,例如离群值的处理和变换变量为近似正态的和同方差的.

现代计算机设备的有效利用使得分析大型数据集成为现实. 这种计算能力促使多元方法应用于图像分析等新领域,以及气象数据等数据的更有效的分析. 同时,一些新的统计分析问题也随之而来,比如参数或数据阵的稀疏性. 由于硬件和软件的迅猛发展以及编译程序需要特殊的知识,在本书中将会对涉及计算的部分给出一些说明. 大多数的方法都有可利用的统计程序软件包.